GEO-News: A Geo-Hierarchy based News Search and Extraction

C.Sudeep Reddy
Lab for Spatial Informatics
International Institute of Information Technology
Gachibowli, Hyderabad - 500 082, India
sudeep@research.iiit.ac.in

Harshita Nanduri
Lab for Spatial Informatics
International Institute of Information Technology
Gachibowli, Hyderabad - 500 082, India
harshi.nanduri@gmail.com

P.D.S.R.Sandeep
Lab for Spatial Informatics
International Institute of Information Technology
Gachibowli, Hyderabad - 500 082, India
sandeep.pdsr@gmail.com

Krishna Chaitanya
Lab for Spatial Informatics
International Institute of Information Technology
Gachibowli, Hyderabad - 500 082, India
krishna.nevali@gmail.com

K.S.Rajan
Lab for Spatial Informatics
International Institute of Information Technology
Gachibowli, Hyderabad - 500 082, India
rajan@iiit.ac.in

ABSTRACT

News is information about recent and important events that have happened at a geographic location at some point in time. Since Location and Time are important components of almost all news, it is important to be able to collate and search news based on time and location and not just keywords. In this paper we present an approach that can deal with not just the geography/location in the query but also consider the Geo-relationship to help pick and rank the most relevant results by introducing a Geo-hierarchy. The results for the NTCIR Geo-Time queries shows a lot of promise for this approach.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval models, search process.

General Terms

Experimentation, Performance, Measurement

Keywords

Geo-Time, Geo-Hierarchy, Information Retrieval, IR evaluation, News

1. INTRODUCTION

News is the communication of information on current and past events which is presented by print, broadcast, Internet, or word of mouth to a third party or mass audience. News, reflect the current happenings and events in various fields like business, politics, Sports, entertainment, Science and technology etc. Web portals provide readers with an option of searching for past news articles. This makes news portals store large volumes of news data. This vast potential for article storage, however, carries its own set of complications. Online searches for newspaper articles often yield a flood of links of least significance. Online searches must yield more useful results to sustain web portal as news reader's primary choice. News publishers have needed better ways of indexing articles, and sought indexing software to make the process easier.

Though Indexing and Information extraction refines the vast number of results, they fail in addressing the Geo-based news results. Consider an article which has "Hyderabad" city in it and discusses about Asian Games. Suppose a query is asked like "In which country are Asian games held?", the pattern matching wont yield any result since the place mentioned in the article is Hyderabad but not India. So, there is a need for a level of Geo-intelligence that can recognize that the location Hyderabad is within the country India and respond accordingly at the higher level of Geo-hierarchy. In order to meet the user's Geo-based query, we present a Geo-hierarchal classification model which displays news in a desired location, if it occurs explicitly and news in its neighbourhood depending on the Geo-relationship. The main objective of this paper is to implement a hierarchical classification of location, with continents as the main category, with countries as its sub category, followed by states, counties and cities.

2. RELATED WORK

Geographic information retrieval is concerned with the retrieval of thematically and geographically relevant information resources in response to a query of the form (theme or topic, spatial relationship, location), "Temples within 5 km. of Tokyo" [6]. Systems that support GIR, such as

geographic digital libraries, and location-aware web search engines, are based on a collection of Geo-referenced information resources and methods to spatially search these resources with geographic location as a key [8, 9].

Information resources are considered Geo-referenced if they are spatially indexed by one or more regions on the surface of the Earth, where the specific locations of these regions are encoded either directly as spatial coordinates, i.e. geometrically, or indirectly by location [7, 9]. However, in order for locations to support a spatial approach to GIR, they must be associated with a model of geographic space. The temporal aspects of search [5] have been largely ignored in the IR community, but not in the GIS and information processing communities.

Existing major news systems display news based on specific keywords or topic wise unlike geographic location of the event. Similarly, spatial news systems matches the location in the queries to the location in articles and the set of news articles presented to readers are static. Though Newsstand [10] handles the Geo-based spatial queries, it is confined to either feature based or location based [16]. Newsstand in spite of presenting readers with Geo-based and feature-based results, lacks the option of providing users with Geo-intelligent responses that can come from the embedded nature of Geo-relationships in the user queries..

GEO-News uses hierarchy and probability functions to decide whether to display articles with keywords match but not just a location match. Current approaches are mainly country based and large numbers of results are expected to be displayed which might be a tedious task for the user to find the articles related to the desired location.

Consider an article with a keyword "Albert Einstein". It may contain "Germany" but doesn't discuss about it. So just a relationship is required to be established between the location and the keywords. Existing works fail to answer this issue. To address this, we follow a process model to implement hierarchal approach of location based searches in this paper. In this paper we focus more on hierarchical relationships of locations rather than on time.

3. ARCHITECTURE AND APPROACH

The main architecture of GEO-News is divided into different modules which are designed to serve a specific purpose. The major modules in the architecture are Preprocessing and Indexing, Pattern matching, and Hierarchical classification.

GEO-News stands out in the aspect of integrating the Geo-data and event data and display the news with refined search results. Major Implementation of the GEO-News includes displaying news within a hierarchal range of a desired location. This can be achieved as explained below:

1. The news data is indexed based on keywords as well as the location using Stanford Named Entity Recognizer (NER) and Stanford Tagger.
2. Keywords (verb, adjective and noun), location and preposition are extracted for the input query using Stanford tools.
3. Pattern matching is implemented to match the keywords in the given query with the index.
4. Among the obtained search results, ones which match the location are assigned a score.

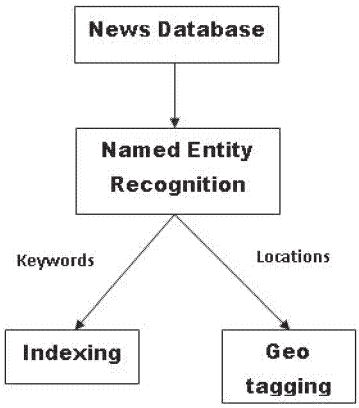

Figure 1: Preprocessing

5. Each preposition has pre-defined hierarchy of 5 levels (continents, countries, states, counties and cities). Construction of hierarchy is implemented with continents occupying the top layer and cities occupying the bottom layer. Based on that hierarchy the movement of the architecture (up/down) is determined and the score is assigned accordingly.
6. Obtained news articles are displayed in based on the rank i.e., descending order of the scores.

The detailed architecture of preprocessing and the query processing are shown in Figure 1 and Figure 2 respectively. Updates to the geographical hierarchy are out of scope of this paper.

3.1 Preprocessing and Indexing

First the keywords (verbs, nouns, and adjectives) are identified from the articles using the statistical natural language processing (NLP) [13]. Stanford indexing system caters to the newspaper industry and servers as an effecting indexing tool for news items. The process of constructing an inverted index is termed as indexing construction or indexing and the machine that performs is the indexer [17]. Blocked sort-based indexing is an efficient single-machine algorithm designed for static collections and single-pass-in-memory indexing, an algorithm with better scaling properties. News portals which contain large collections of data, indexing has to be distributed over computer clusters with hundreds or thousands of machines. Collections are often so large that we cannot perform index construction efficiently on a single machine. This is particularly true of the World Wide Web for which we need large computer clusters to construct any reasonably sized web index. Web search engines, therefore, use distributed indexing algorithms for index construction [2, 14]. The result of the construction process is a distributed index that is partitioned across several machines - either according to term or according to document.

Collections with frequent changes require dynamic indexing [15] so that changes in the collection are immediately reflected in the index. Identifying entities from different language news articles is another major issue to be addressed. A pre-defined set of linguistic features are designed to identify named entities which are language independent [1, 12].

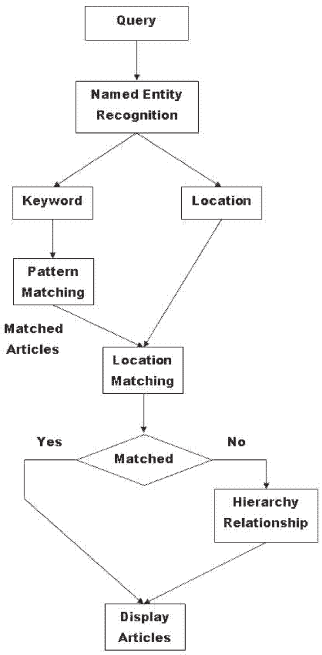

Figure 2: Query Processing

We use, Stanford Part of Speech (POS) tagger [18] implementing the distributed and dynamic indexing which serves as an efficient indexing tool for the news articles. In addition, Stanford tagger reads text in some language and assigns part of speech to each keyword such as noun, adjective, verb, etc., although generally computational applications use more fine-grained POS tags like "noun-plural". This software is a Java implementation of the log-linear part-of-speech taggers [11, 18]. Indexing a subset of information extraction, assigns index for keywords (noun, adjective, verb, etc.) and location.

GEO-News uses Named-Entity Recognition (NER), a subtask of Information Extraction is a well-studied problem in Natural Language Processing (NLP) which is concerned with identifying entities such as person, location, and organization names. Previous NER systems are built on rule based approaches that use manually built finite state patterns matching a sequence of words in the same manner as matching a regular expression. Rule based approaches failed to address the robustness and portability [4]. In Geo-News we use a machine learning approach (statistical approach) which is attractive, adaptable, and trainable and a cheaper maintenance approach than rule-based one. In the statistical approach, training data must be matched with the entities of interest and their types. We use a Stanford Named Entity Recognition (NER) which labels sequences of words in a text which are the names of things, such as person and company names, or sports or business names. The software provides a general (arbitrary order) implementation of linear chain Conditional Random Field (CRF) sequence models [3], coupled with well-engineered feature extractors for Named Entity Recognition. The distributional similarity features improve performance but the models require considerably more memory.

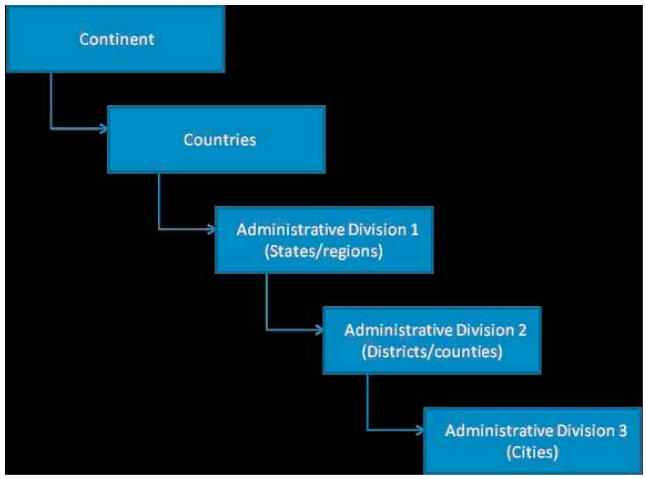

Figure 3: Model of the hierarchy

3.1.1 Geo-Tagging

Geo-Tagging addresses the Geo-based constraint using sources like Gazetteers (e.g. GeoNames) which contain locations in one or more languages, latitude and longitude information as well as hierarchical information, indicating that a town belongs to a county, which is part of a region, which is itself (region) part of a country, etc.

Geo-Tagging model needs to address the non local feature structure for displaying articles which are non local (near to the desired city, within state, country and continent). Most statistical models currently used in natural language processing represent only local structure. Although this constraint is critical in enabling tractable model inference, it is a key limitation in many tasks, since natural language contains a great deal of nonlocal structure. A general method for solving this problem is to relax the requirement of exact inference, substituting approximate inference algorithms instead, thereby permitting tractable inference in models with non-local structure. Gibbs sampling, a simple Monte Carlo algorithm that is appropriate for inference in any factored probabilistic model, including sequence models and probabilistic context free grammars. Although Gibbs sampling is widely used elsewhere, there has been extremely little use of it in natural language processing. Here, we use it to add non-local dependencies to sequence models for information extraction [13]. A constraint model can be effectively combined with an existing sequence model in a factored architecture to successfully impose various sorts of long distance constraints.

Building a hierarchy of geographical locations is a language independent task. There are seven continents in the world and a total of 195 countries distributed in these seven continents. There will be many sub-divisions within a country. In the proposed hierarchy we have "Continents" as the main category, in which we have "Countries" in the corresponding continents as the sub category, in which we have "first order administrative divisions" (for instance states in India) in the country as the third level category and "second order administrative divisions" of the country (for instance districts in states of India) as the fourth level category and "third order administrative divisions" (Cities) as the fifth level category.

In Figure 3 "Administrative Division 2" is optional as many countries like "Singapore", "Qatar" etc won't have second order sub divisions. In New York Times data we can get the information about countries in a continent, first order sub-divisions (states) in a country and second order subdivisions (districts) in a country etc. But this information is not in a common format (unstructured) for all the countries. So it was difficult to collect this geographical information from New York Times data. So we used external sources like "Gazetteers" for building the hierarchy.

Gazetteers are geographical dictionaries which contain information like locations, latitude, longitude information as well as hierarchical information, indicating that a town belongs to a county, which is part of a region, which itself (region) is part of a country.

We considered gazetteers like GeoNames[1], KNAB database[2], Alexandria gazetteer[3] for building the hierarchy. But "GeoNames" has covered more number of locations (geographical locations) compared to other gazetteers. So we used "GeoNames" for construction of the hierarchy.

3.1.2 *Construction of the hierarchy*

1. GeoNames database consists of a table (named CountryInfo) which has the information about the country and its corresponding continent. We used the information in this table to construct up to the second level (countries) of the hierarchy.

2. We had extracted all the entries in the main "geoname" table (allCountries), containing "feature code" as "ADM1". With this information we had constructed the third level of the hierarchy.

3. Then we had extracted all the entries that have "ADM2" in the "geoname" table (allCountries) as the feature code and constructed up to the fourth level.

4. We faced some problems in constructing the fifth level of the hierarchy. This is because the entry of city in "allCountries" table had no information about the second order administrative division to which this city belongs to. Due to this we are unable to find a link between "second order division" and "third order division (cities)" of a country. Therefore, we went for another external resource which contained the information (http://www.world gazetteer.com/dataen.zip) about a location and the divisions (like country -> state -> district) above it. We built the fifth level of the hierarchy using this external resource. But it has covered less number of locations compared to the locations in the GeoNames database.

We can see in Figure 4 of hierarchy, there is no second order administrative division in Singapore. There is a direct link between first order administrative division and cities in Singapore. One solution for this problem is completely removing the second order administrative division of a country in the hierarchy and having a direct link between the first order administrative division and cities (third order administrative division). Another solution is obtaining this connection by using sources like ISD codes of countries; STD codes of cities in those countries or by using the "allCountries" geonames table.

[1]http://download.geonames.org/export/dump
[2]http://www.eki.ee/knab/p_mm_en.htm
[3]http://www.alexandria.ucsb.edu/gazetteer/

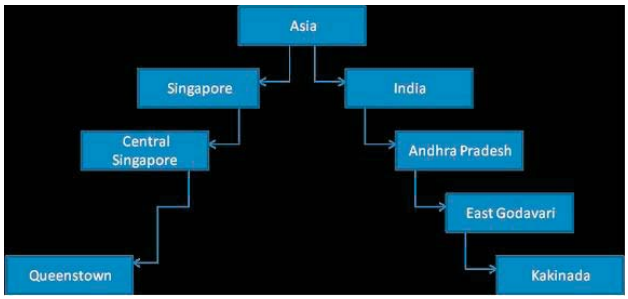

Figure 4: Part of Asia continent in the constructed hierarchy

Now the New York Times articles need to be tagged to a specific geographical location. Stanford Named Entity Recognizer (NER) is used to obtain names of locations. Two approaches are used for tagging the Geographic location to the related article.

Approach 1.

By considering only the title and information box of the article:

1. Check whether "title" is a continent or not. If it is not a continent then check the database. Ex: New York Times article of Asia can be uniquely tagged to the continent "Asia".

2. On checking if we are unable to disambiguate it (i.e., if we are unable to find the exact location to which this article refers to) then use the information box (if present, else check the content of the article for disambiguation) to know whether the location can be uniquely identified or not. Ex: Consider New York Times article of "London", we cannot disambiguate it using the database because U.S.A alone contains more than 10 locations named London. So if we use information box present in the "London" article, there will be attributes like "Sovereign state: United Kingdom", "Constituent Country: England", "Region: London" and "Districts: City" and 32 boroughs. These attributes help in identifying the exact location to which this article refers.

3. If information box haven't uniquely identified the location (i.e., if there isn't enough information about the location in the info box) then check the content of the article.

Approach 2.

By considering the content of the article:

Apply named entity recognition algorithm (used Stanford Named Entity Recognizer) to obtain the geographic names in the article. Next we have taken all the geographic names (those which are marked as LOCATION by the NER) mentioned in a sentence as a group. So a set of groups which contains all the geographic names mentioned in the page, is obtained.

3.2 Pattern matching

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. An object is an entity which keeps or stores information in a database. User queries are matched to objects stored in the database. Most IR systems compute a numeric score on how well each object in the database match the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query.

This phase involves matching the keywords from the query with the indexed keywords. Based on the number of matches score is assigned like 3 keywords matched, 2 keywords matches, 1 keyword matched etc. Then the obtained articles are checked for location matches. If the location matches with the location in the index, a score is assigned to it. Now we have the articles for which there is keyword match (atleast one) and location not matching left behind. We use hierarchy and probability functions here to decide whether to display them are not. Prepositions are used for this purpose.

3.3 Hierarchy classification

Existing news query systems display limited number of results for a location matching news articles. These results are confined to the desired location in the query and might fail to meet the user's requirements in displaying the necessary articles. Hierarchal classification overcomes this limitation by providing user with the location match articles and articles which fall under either near by city or same state, country and continent, there by displaying more results which results in higher probability of meeting the user's requirement.

Now we have to disambiguate (assigning a standard unique taxonomy like A -> B -> C to each phrase in the text, which is considered to be a location) the geographic names as there might be more than one location having the same geographic name. Rules for disambiguation are:

1. If the geographic names in a group can uniquely qualify a spot (another geographic name in the same group), then assign a unique meaning to the spot with a score in the range of 0.95 - 1, to indicate its high level of certainty. Ex: Consider Nalgonda, Andhra Pradesh, India obtained from "Nagarjuna Sagar is located in Nalgonda, Andhra Pradesh, India". In the database we can find a unique entry for this set. So assign a unique meaning (taxonomy) to "Nalgonda" like Nalgonda -> Andhra Pradesh -> India -> Asia. Combinations which are not unique like Hyderabad, Asia (of which there are many) or Paris, Amalapuram (of which there are none i.e., no entry in the database) are not assigned any score.
2. If there are multiple spots with the same geographic name occurring in the page, of which only one is qualified (unique with score in range of 0.95 - 1) then taxonomy of the qualified spot is assigned to the other spots with a score in the range of 0.8 - 0.9.
3. Now consider all the spots which have score below 0.6 and check the common locations in their taxonomies. Ex: Consider London which is present in "England, United Kingdom", "Ontario, Canada", "California, U.S.A" and Hamilton which is present in "Victoria, Australia", "Ontario, Canada" etc. Both of these share "Ontario, Canada". Therefore London is assigned to London -> Ontario -> Canada and Hamilton is assigned to Hamilton -> Ontario -> Canada. A score in the range of 0.65 to 0.75 is assigned to them. This is due to the reason that the writer of the article may haven't specified the clear meaning thinking that it is implied from the context. So we can assume that all the geographic names share a context.

Now we have geographic names each with a unique taxonomy and a score associated with them. Now using these scores find the geographic location of the article. The geographic location can be a city, county, state, country or a continent. Use the scores of the locations such that for A -> B -> C, importance of A should be more than B and importance of B to be more than C. If the computation is done in this way the lower levels in the taxonomy gets higher preference so that location of the article can be identified up to the deepest level. Ex: Tagging "Thousand Pillar Temple" should give output as Hanumakonda (Hanumakonda -> Andhra Pradesh -> India -> Asia) and not Andhra Pradesh.

3.3.1 Algorithm

Take the disambiguated geographical locations in the query. Initialize the Fscore (final score) of all geographical locations to zero.

Do for each location left for computation

/* Let "A -> B -> C" be taxonomy of the location and its score be 'p'. */

Step 1.

Add p^2 to the Fscore of A -> B -> C.

Step 2.

Add $(p^2) * d$ to Fscore of B -> C. // d (decrement) is a constant.

Step 3.

Add $(p^2) * (d^2)$ to Fscore of C.

Quadratic function is used in order to increase the relative weight of the disambiguations with higher score.

Now sort the resulting taxonomy levels by Fscore. Loop from the highest to the lowest Fscore taxonomies stopping at a threshold c. We can select more than one location as the article's location. We won't consider taxonomy levels that are covered by already selected location for the article (Eg: Andhra Pradesh won't be selected if Guntur -> Andhra Pradesh -> ... is already selected).

Example.

Let a article has two mentions of Nalgonda -> Andhra Pradesh -> India -> Asia (Fscore = 0.85), two Guntur -> Andhra Pradesh -> India -> Asia (Fscore = 0.85), three Hyderabad -> Andhra Pradesh -> India -> Asia (0.67) and one Asia (0.97). Let 'd' (decrement) be 0.6

Fscore(Nalgonda -> ...) $= 2 * ((0.85)^2) = 1.44$

Fscore(Guntur -> ...) $= 2 * ((0.85)^2) = 1.44$

Fscore(Andhra Pradesh -> ...) $= [2 * (0.85^2) * (0.6) + 2 * (0.85^2) * (0.6) + 3 * (0.67^2) * (0.6)] = 2.54$

Table 1: Precision and Recall values of the retrieved articles

Query ID	Recall	Precision
GeoTime-0001	0.7	1
GeoTime-0002	0.98	1
GeoTime-0003	0.62	1
GeoTime-0004	0.76	1
GeoTime-0005	0.95	1
GeoTime-0006	0.85	1
GeoTime-0007	0.75	1
GeoTime-0008	0.79	1
GeoTime-0009	0.87	1
GeoTime-0010	0.87	1
GeoTime-0011	0.83	1
GeoTime-0012	0.96	1
GeoTime-0013	0.9	1
GeoTime-0014	0.83	1
GeoTime-0015	0.9	1
GeoTime-0016	0.95	1
GeoTime-0017	0.94	1
GeoTime-0018	0.89	1
GeoTime-0019	0.82	1
GeoTime-0020	0.78	1
GeoTime-0021	0.43	1
GeoTime-0022	0.84	1
GeoTime-0023	0.83	1
GeoTime-0024	0.84	1
GeoTime-0025	0.74	1

Fscore(India) = $[2*(0.85^2)*(0.6^2)+2*(0.85^2)*(0.6^2)+3*(0.67^2)*(0.6^2)] = 1.52$

Fscore(Asia) = $[2*(0.85^2)*(0.6^3)+2*(0.85^2)*(0.6^3)+3*(0.67^2)*(0.6^3)] = 0.915$

Fscore(Hyderabad) = $3*(0.67^2) = 1.346$

If we choose threshold level as 1 then the candidates that can be the focus of the article are Andhra Pradesh, India, Nalgonda, Guntur, Hyderabad. India is removed as Andhra Pradesh has already covered it. So the foci are Andhra Pradesh, Nalgonda, Guntur and Hyderabad. We can consider only Nalgonda, Guntur and Hyderabad as they all have Andhra Pradesh in the taxonomy.

Having multiple locations to a article is useful because if we take "Nagarjuna sagar" which is 20 kms from Guntur and also nearer to Nalgonda. So we can tag the article to both Guntur and Nalgonda, if we have more focus for the article.

4. EXPERIMENTAL ANALYSIS

4.1 Data Sets

News story collections from New York Times is used for evaluation. The English collection consisted of 315,417 New York Times stories also for 2002-2005. Since we were interested in looking for particular events around which geotemporal topics could be constructed, we ran frequency distributions on both collections by month and discovered gaps in the NYT collection for Jan 2003-July 2004. While the monthly average of documents for 2002 and 2005 was 9,982 and 8,703 respectively, for 2003 and 2004 it was 2,319 and 5,280. Indeed from January 2003 through June 2004 (zero documents), the number of documents per month ranged from 0 to 2209 documents (see the GeoTime collection page http://metadata.berkeley.edu/NTCIR-GeoTime/databases.php for the complete distributions).

4.2 Experimental Evaluation

The performance of the GEO-News is tested on the 25 queries given by NTCIR group. Considering the results obtained, we found out the precision and recall of all the results we obtained. In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the search and recall is the fraction of the documents that are relevant to the query that are successfully retrieved. We have used lucene for information retrieval. The precision and recall values of the retrieved articles for all the given 25 articles can be seen in Table 1.

The recall value of the query results varies between 0.43 to 0.98. The low values are due to stemming issues as the stemming and parsing results are obtained by using Stanford Parser, which is not perfect. The average recall value is 0.8208. Query 2 (GeoTime-0002), "When and where did Hurricane Katrina make landfall in the United States?", gave the best recall value as the query involved key word based search and used Geo-hierarchy. Moreover, this query did not have many words which need to be stemmed and since we had to go down the hierarchy to find the locations where hurricane katrina made a landfall. Query 21 (GeoTime-0021),"When and where were the 2010 Winter Olympics host city location announced?", gave the least recall value as this is not a pure keyword based query. It involves lots of words to be stemmed and the Stanford Parser did not do this efficiently which resulted in giving less accurate results.

We can observe here that the recall value of GeoTime-0002 is more than that of GeoTime-0001, as in that query Geo-hierarchy is being applied, hence giving better results. Where as, few queries give less recall value due to problems in stemming. Since all the retrieved documents are relevant to the query the precision is always 1.

5. CONCLUSIONS

In this paper the articles are indexed based on keywords as well as the geographic location and then Geo-tagged. While pattern matching, instead of just keyword matching, location is also considered and the score is assigned based on the hierarchy of the geographic location. This results in the more accurate results, which are of user's interest. The Hierarchical relationships answer the queries more precisely and accurately. Narration of query is used as well, which helps in the finding a where part thereby improving the accuracy of the results.

This work can be extended to handle range queries such as "articles around 50km from Hyderabad". Period questions like "articles occured in 2 years" needs to be handles. Also the task of tagging the articles with geographical locations, to multiple languages needs to be done.

6. ACKNOWLEDGMENTS

We would like to thank Balaji from IR lab for helping us in gathering the papers related to Information Retrieval tools and Padmini Priyadarshini for giving valuable suggestions in writing the paper. We would like to thank NTCIR-8 organizers for introducing the Geo-Time challenge and Linguistic

Data Consortium providing the collections for research.

7. REFERENCES

[1] Asif Ekbal et. al, Language Independent Named Entity Recognition in Indian Languages, IJCNLP, 2008.

[2] C. B. Jones, A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT spatial search engine: architecture, ontologies and spatial indexing, GiScience 2004, Adelphi, MD, pages 125Ű139, 2004.

[3] D Santos, L Cabral, GikiCLEF: Crosscultural issues in an international setting: asking non-English-centered questions to Wikipedia, CLEF 2009 Working Notes, $http://www.clef-campaign.org/2009/working_notes/Santos-paperCLEF2009.pdf$, September 2009, 21pp.

[4] Finkel J. R., Grenager T., and Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling, In Proceedings of the 43rd Annual Meeting on Association For Computational Linguistics (Ann Arbor, Michigan, June 25 - 30, 2005). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 363-370. DOI = http://dx.doi.org/10.3115/1219840.1219885

[5] I. Mani, J. Pustejovsky, and B. Sundheim. Introduction to the special issue on temporal information processing. ACM Transactions on Asian Language Information Processing (TALIP), 3(1):1Ű10, 2004.

[6] Larson, R Geographic information retrieval and spatial browsing, In GIS and Libraries: Patrons, Maps and Spatial Information, pages 81Ű124, UIUC - GSLIS, Urbana-Champaign, IL, 1996.

[7] L L Hill, GeoReferencing: The Geographic Associations of Information, MIT Press, Cambridge, MA 2006.

[8] R. Purves, C. Jones, and P. Clough. GIRŠ10: 6th workshop on geographic information retrieval, 2010. http://www.geo.unizh.ch/rsp/gir10/index.html

[9] S Asadi, C.-Y. Chang, X. Zhou, and J. Diederich, Searching the world wide web for local services and facilities: A review on the patterns of location-based queries, pages 91Ű101, WAIM2005, Springer LNCS 3739, 2005.

[10] Teitler B. E., Lieberman M. D., Panozzo D., Sankaranarayanan J., Samet H., and Sperling J. 2008, NewsStand: a new view on news, In Proceedings of the 16th ACM SIGSPATIAL international Conference on Advances in Geographic information Systems (Irvine, California, November 05 - 07, 2008), GIS '08. ACM, New York. DOI = http://doi.acm.org/10.1145/1463434.1463458

[11] Toutanova K., and Manning C. D., Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association For Computational Linguistics - Volume 13 (Hong Kong, October 07 - 08, 2000), Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 63-70. DOI = http://dx.doi.org/10.3115/1117794.1117802

[12] Zhang L., Pan Y., and Zhang T., Focused named entity recognition using machine learning, In Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004), SIGIR '04 ACM, New York, NY, 281-288. DOI = http://doi.acm.org/10.1145/1008992.1009042

[13] Zhou G., and Su J., Named entity recognition using an HMM-based chunk tagger, In Proceedings of the 40th Annual Meeting on Association For Computational Linguistics (Philadelphia, Pennsylvania, July 07 - 12, 2002), Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 473-480. DOI = http://dx.doi.org/10.3115/1073083.1073163

[14] $http://nlp.stanford.edu/IR-book/html/htmledition/distributed-indexing-1.html\#sec:distributedindexing$

[15] $http://nlp.stanford.edu/IR-book/html/htmledition/dynamic-indexing1.html\#sec:dynamicindexing$

[16] http://www.metacarta.com/

[17] http://nlp.stanford.edu/IR-book/html/htmledition/index-construction-1.html

[18] http://nlp.stanford.edu/software/tagger.shtml